[Data Mining Introduction – Different Types of Data - Technologies Used - Issues in Data Mining – Applications – Data Objects and Attributes - Google Slides]

Data Preprocessing:

1. Data Preprocessing: An Overview

Data preprocessing is a crucial step in the data mining process that aims to prepare the data for analysis. It involves several tasks that help improve the quality of the data, making it more suitable for data mining algorithms. The primary tasks include data cleaning, integration, reduction, and transformation.

2. Data Quality: Why Preprocess the Data?

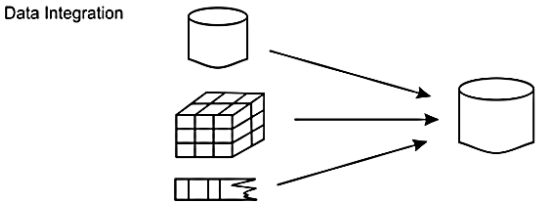
Data quality directly affects the accuracy of data mining results. Several factors define the quality of data:

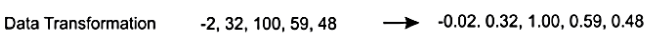
- **Accuracy**: Ensures data is correct and precise.
- **Completeness**: Addresses missing or unavailable data.
- **Consistency**: Ensures no discrepancies or contradictions.
- **Timeliness**: Ensures data is updated and current.
- **Believability**: Evaluates the trustworthiness of data.
- **Interpretability**: Ensures data is easy to understand.

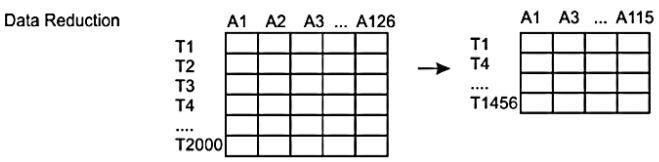
3. Major Tasks in Data Preprocessing

The major tasks involved in data preprocessing include:

- **Data Cleaning**: Involves handling missing values, correcting noisy data, detecting and handling outliers, and resolving inconsistencies in the data.
- **Data Integration**: Combining data from multiple sources into a cohesive dataset, ensuring schema integration and resolving conflicts.
- **Data Reduction**: Reducing the volume of data while preserving the quality and analytical results.
- **Data Transformation**: Modifying data into suitable formats for analysis (e.g., normalization, discretization).







Data Cleaning

1. Data in the Real World is Dirty

Real-world data often contains inaccuracies such as missing, noisy, or inconsistent values. These errors can arise from various sources, including faulty instruments, human mistakes, or data entry errors.

- **Incomplete Data**: Missing values for certain attributes or entries.
- **Noisy Data**: Data with errors, such as wrong or out-of-range values.
- **Inconsistent Data**: Data with conflicting values or formats (e.g., "Age" conflicting with "Birthday").

2. Handling Missing Data

- **Ignore the Tuple**: Not recommended when missing data is extensive.
- **Manual Filling**: Tedious and impractical for large datasets.
- **Automatic Filling**: Methods include using global constants, mean values, or inference-based approaches (e.g., using decision trees or Bayesian methods).

3. Handling Noisy Data

- **Binning**: Data is sorted and divided into bins, and values are smoothed by taking the average or median within each bin.
- **Regression**: Fitting data to regression models to smooth out errors.
- **Clustering**: Identifying and removing outliers by grouping similar data points together.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Data Integration

1. Schema Integration

Data integration involves combining data from different sources into a coherent dataset. This includes schema integration, where attributes like "customer_id" in one source may be represented differently in another (e.g., "cust_id" vs. "cust_no").

2. Handling Redundancy

Handling Redundancy in Data Integration

Redundant data often occurs when integrating multiple databases. Common issues include:

- **Object Identification**: Same attribute or object with different names.
- **Derivable Data**: One attribute derived from another (e.g., annual revenue).
- **Redundant Attributes**: Detected through correlation and covariance analysis.

Correlation Analysis (Nominal Data)

- **Chi-Square (X^2) Test**: Used to assess the relationship between two variables:

$$\chi^2 = \sum \left(\frac{(Observed - Expected)^2}{Expected} \right)$$

- **Interpretation**: A higher X^2 value indicates a stronger relationship, but correlation does not imply causality.

Chi-Square Calculation Example:

	Play Chess	Not Play Chess	Sum (row)
Like Science Fiction	250 (90)	50 (210)	200
Not Like Science Fiction	200 (360)	1000 (840)	450
Sum (col.)	450	1050	1500

- **Chi-Square Calculation**:

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- The result shows that the variables "like science fiction" and "play chess" are correlated in the group.

Degree of Freedom (df)

For a contingency table with r rows and c columns, the degree of freedom is calculated as:

$$df = (r-1) \times (c-1)$$

Here, the table has 2 rows (like and not like science fiction) and 2 columns (play chess and not play chess), so:

$$df = (2-1) \times (2-1) = 1$$

Conclusion

- **Chi-Square Value**: 507.93
- **Degree of Freedom**: 1

Next, we compare the calculated chi-square value (507.93) with the critical value from the chi-square distribution table for $df = 1$ at a significance level (typically $\alpha = 0.05$).

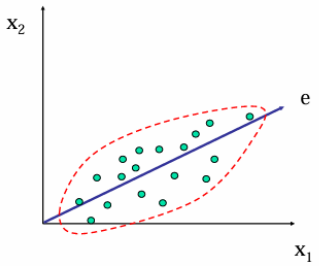
- The critical value for $df = 1$ at $\alpha = 0.05$ is 3.841.

Since the calculated chi-square value (507.93) is much greater than the critical value (3.841), we **reject the null hypothesis**. This indicates that there is a statistically significant relationship between "like science fiction" and "play chess."

Thus, **"like science fiction" and "play chess" are correlated** in this group, meaning the two variables are related. However, correlation does not imply causality.

Data Reduction

1. Dimensionality Reduction
Reducing the number of attributes (features) in a dataset without losing essential information. Techniques include:

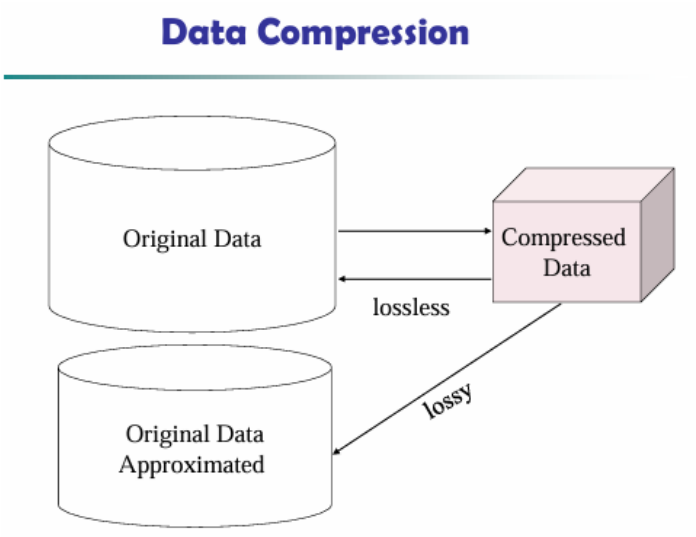


- **Principal Component Analysis (PCA)**: Reduces dimensionality by projecting data onto a smaller space that captures the most variance.
- **Feature Subset Selection**: Identifying and selecting relevant features while discarding irrelevant or redundant ones.

2. Numerosity Reduction
Reducing the volume of data by using alternative, smaller forms of data representation. Methods include:

- **Regression and Log-Linear Models**: Fit data to a model, estimating parameters and discarding raw data.
- **Histograms, Clustering, Sampling**: Using summarization techniques like histograms and clustering to represent large datasets with fewer data points.

3. Data Compression
Compression techniques reduce data size while retaining important information. This includes:



- **Lossless Compression**: Techniques like string compression, which allow full reconstruction of the original data.
- **Lossy Compression**: Methods like audio and video compression, where some information is discarded for efficiency.

Example: In **predicting flat prices**, the dataset might include multiple features like location, square footage, number of rooms, age of the building, and amenities. By applying dimensionality reduction techniques like PCA, we can reduce the number of features, focusing on the most relevant ones (e.g., location and square footage), thus improving the efficiency of the prediction model.

Data Transformation

Data transformation involves changing or mapping attribute values to a new set, ensuring each old value matches a new one. Key methods include:

1. **Smoothing**: Removes noise from data.
2. **Attribute Construction**: Creates new attributes from existing ones.
3. **Aggregation**: Summarizes data or creates data cubes.
4. **Normalization**: Scales data within a range (e.g., min-max, Z-score, or decimal scaling).
5. **Discretization**: Divides continuous data into intervals for simplification.

SAND

Normalization Methods:

- **Min-Max Normalization**: Scales values to a specified range.
- **Z-Score Normalization**: Scales based on mean and standard deviation.
- **Decimal Scaling**: Shifts the decimal point to scale data.

Data Set: **200, 300, 400, 600, 1000**

1. Min-Max Normalization

Formula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where:

- $X_{\min} = 200$ (minimum value)
- $X_{\max} = 1000$ (maximum value)

Min-Max Normalized values:

- For 200:
$$X' = \frac{200 - 200}{1000 - 200} = 0$$
- For 300:
$$X' = \frac{300 - 200}{1000 - 200} = 0.125$$
- For 400:
$$X' = \frac{400 - 200}{1000 - 200} = 0.25$$
- For 600:
$$X' = \frac{600 - 200}{1000 - 200} = 0.5$$
- For 1000:
$$X' = \frac{1000 - 200}{1000 - 200} = 1$$

Final Min-Max Normalized values: **0, 0.125, 0.25, 0.5, 1**

2. Z-Score Normalization

Formula:

$$X' = \frac{X - \mu}{\sigma}$$

Where:

- μ = Mean of dataset
- σ = Standard deviation of dataset

Step 1: Calculate the Mean (μ)

$$\mu = \frac{200 + 300 + 400 + 600 + 1000}{5} = \frac{2500}{5} = 500$$

Step 2: Calculate the Standard Deviation (σ)

$$\sigma = \sqrt{\frac{(200 - 500)^2 + (300 - 500)^2 + (400 - 500)^2 + (600 - 500)^2 + (1000 - 500)^2}{5}}$$

Let's calculate the squared differences:

- $(200 - 500)^2 = 300^2 = 90000$
- $(300 - 500)^2 = 200^2 = 40000$
- $(400 - 500)^2 = 100^2 = 10000$
- $(600 - 500)^2 = 100^2 = 10000$
- $(1000 - 500)^2 = 500^2 = 250000$

Now, sum the squared differences:

$$90000 + 40000 + 10000 + 10000 + 250000 = 400000$$

Next, divide by the number of data points:

$$\frac{400000}{5} = 80000$$

Finally, take the square root:

$$\sigma = \sqrt{80000} \approx 282.84$$

Step 3: Apply the Z-Score formula to each value:

- For 200:
$$X' = \frac{200 - 500}{282.84} \approx -1.06$$
- For 300:
$$X' = \frac{300 - 500}{282.84} \approx -0.71$$
- For 400:
$$X' = \frac{400 - 500}{282.84} \approx -0.35$$
- For 600:
$$X' = \frac{600 - 500}{282.84} \approx 0.35$$
- For 1000:
$$X' = \frac{1000 - 500}{282.84} \approx 1.77$$

Final Z-Score Normalized values: -1.06, -0.71, -0.35, 0.35, 1.77

3. Z-Score Absolute Deviation Normalization (MAD)

Formula:

$$X' = \frac{|X - \mu|}{A}$$

Where:

- A is the Mean Absolute Deviation (MAD), calculated as:

$$A = \frac{\sum |X - \mu|}{N}$$

Step 1: Calculate the Absolute Deviations:

- $|200 - 500| = 300$
- $|300 - 500| = 200$
- $|400 - 500| = 100$
- $|600 - 500| = 100$
- $|1000 - 500| = 500$

Step 2: Calculate the Mean Absolute Deviation (MAD):

Sum of absolute deviations:

$$300 + 200 + 100 + 100 + 500 = 1200$$

MAD:

$$A = \frac{1200}{5} = 240$$

Step 3: Apply the formula for Z-Score Absolute Deviation normalization to each value:

- For 200:

$$X' = \frac{|200 - 500|}{240} = \frac{300}{240} = 1.25$$

- For 300:

$$X' = \frac{|300 - 500|}{240} = \frac{200}{240} = 0.83$$

- For 400:

$$X' = \frac{|400 - 500|}{240} = \frac{100}{240} = 0.42$$

- For 600:

$$X' = \frac{|600 - 500|}{240} = \frac{100}{240} = 0.42$$

- For 1000:

$$X' = \frac{|1000 - 500|}{240} = \frac{500}{240} \approx 2.08$$

Final Z-Score Absolute Deviation values: **1.25**, **0.83**, **0.42**, **0.42**, **2.08**

Summary of Results:

1. Min-Max Normalized values: **0**, **0.125**, **0.25**, **0.5**, **1**
2. Z-Score Normalized values: -1.06, -0.71, -0.35, 0.35, 1.77
3. Z-Score Absolute Deviation values: **1.25**, **0.83**, **0.42**, **0.42**, **2.08**